IX International conference
"Information Technology and Implementation" (IT&I-2022)
Kyiv, Ukraine

Towards the Enterprise Architecture Web Mining Approach and Software Tool

Andrii Kopp and Dmytro Orlovskyi

Research Relevance

What is the **Enterprise Architecture (EA)**?

- the organizing logic for business processes and information technology (IT) infrastructure that reflects integration and standardization requirements of the operating model used by an enterprise;
- the conceptual blueprint that defines the structure and operations of a company, and determines how it can achieve its ongoing and future goals in the most efficient way.

EA could be considered as ***a structured high-level description of an organization*** from different viewpoints (i.e. business, data, applications, and technology) that serve each other.

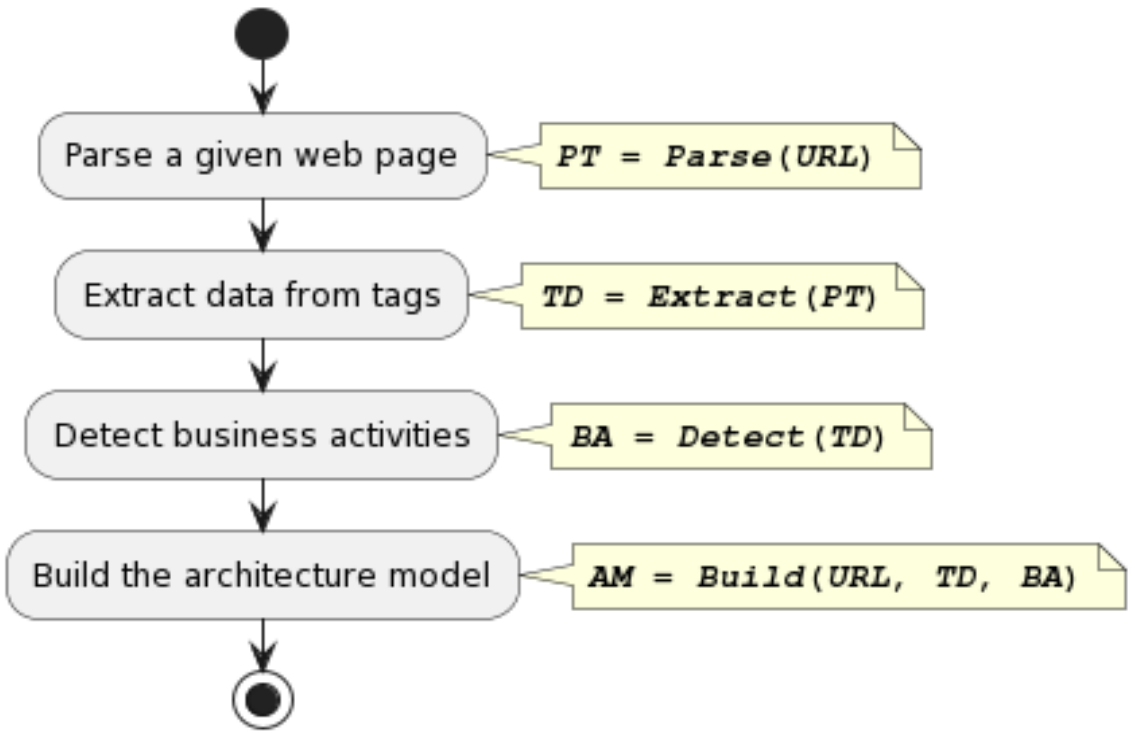
This paper proposes an approach and a software tool for the **automatic extraction of EA landscapes from websites** that nowadays virtually represent organizations on the Internet.
This approach aims at **simplifying the procedure of building high-level models** in the preliminary stages of EA development.

- organizations can **understand the efficiency** of current EA state
- **make decisions** on necessary changes to improve efficiency
- **define gaps** between the ongoing and desired states
- **define initiatives** that should be implemented

Enterprise Architecture Web Mining

The **"EA Web Mining" technique** is focused on the automatic construction of EA models using corporate websites as sources of data about EA elements and the relationships between them.

The main problem is ***finding mentions of business processes and other EA elements*** in HyperText Markup Language (HTML) pages of corporate websites.



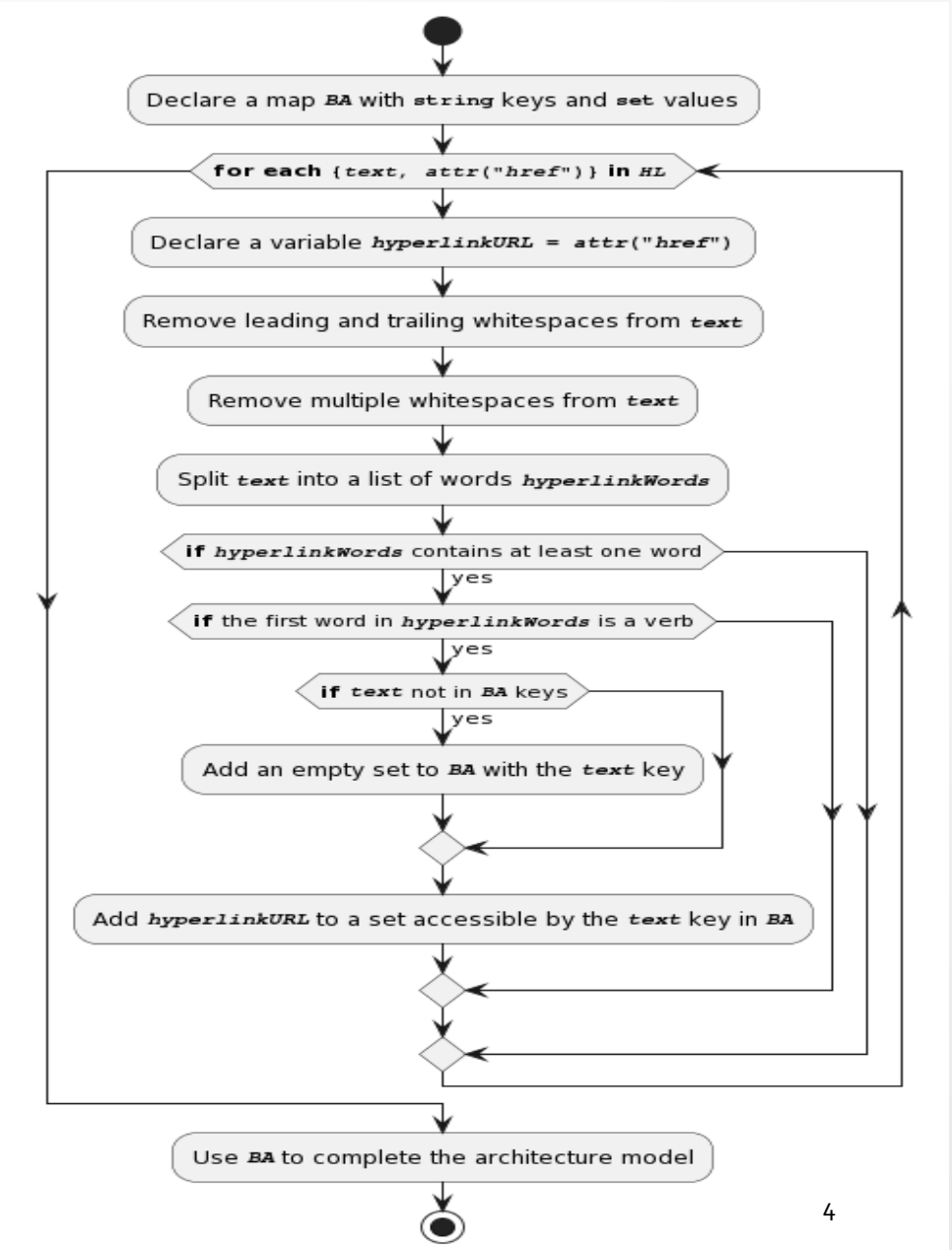
EA models automatically produced using the company's website can help to understand the current state of the enterprise, including its customer relationship strategy, offered products, and services. Then, shortcomings could be detected in such an EA model, and the decisions to **improve the enterprise's virtual representation on the Internet** could be made.

Business Activities Detection in Organizational Web Pages

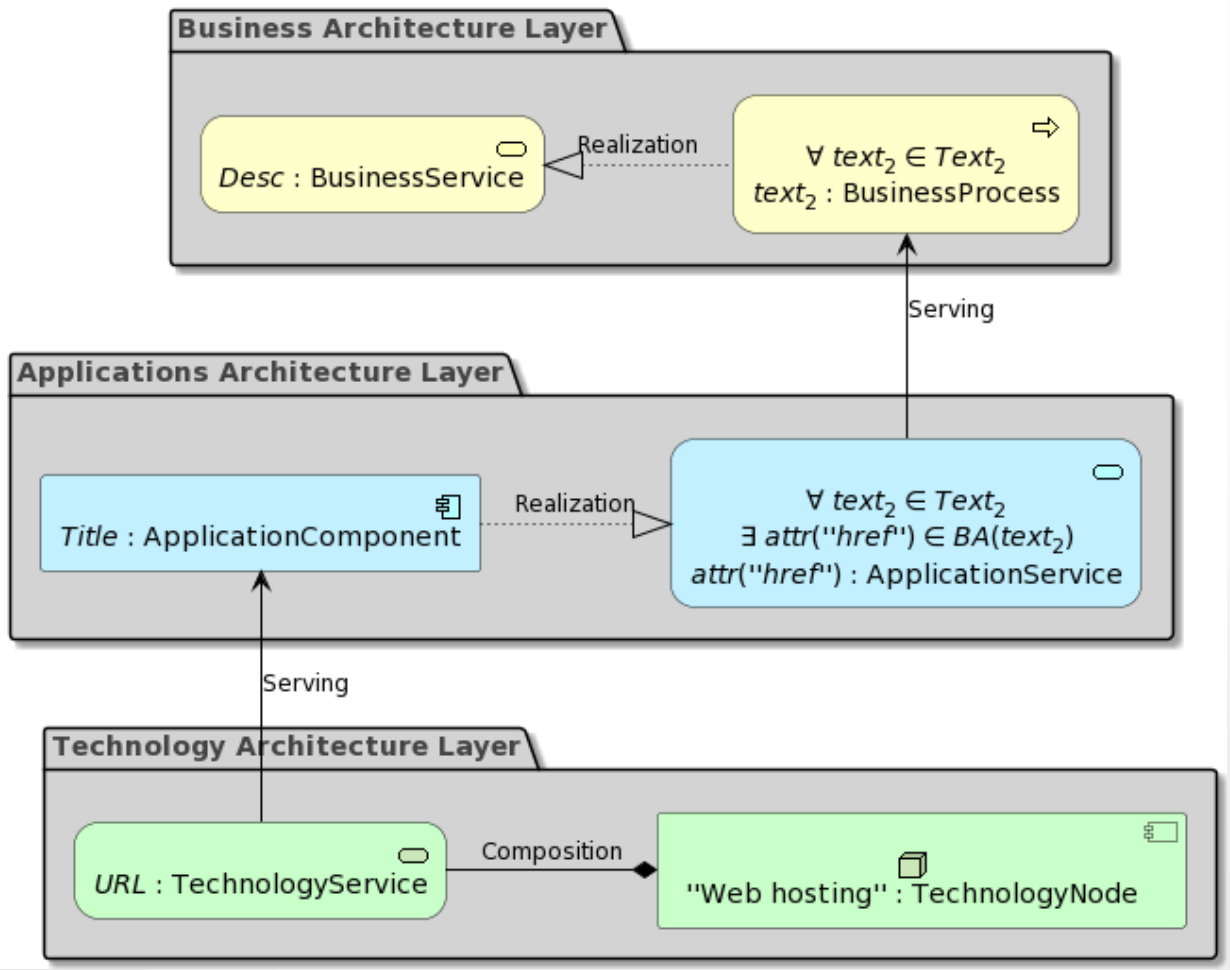
- The **web page description** reflects products or services virtually offered by the organization via its homepage.
- The **web page title** represents the software (website) that supports business processes of products or services delivery through the Internet.
- **Hyperlinks** reflect actions that customers can do when visiting a website to trigger business processes to get products or services.

Business activities detection algorithm:

1. Remove leading and trailing whitespaces from the current hyperlink text content value.
2. Remove multiple whitespaces from the current hyperlink text content value.
3. Split the current hyperlink text value into a bag of words.
4. Tag each word in the bag of words as a part of a speech.
5. If the bag contains at least one word and its first word is a verb, then this hyperlink represents a business activity.



Enterprise Architecture Landscape Construction using ArchiMate



Here:

- $Desc$ is the web page description data;
- $Text_2$ is the set of hyperlink text content values $text_2 \in Text_2$;
- BA is the mapping between hyperlink text content values $text_2 \in Text_2$ and their URL values;
- $attr("href")$ is the hyperlink that correspond to the $text_2 \in Text_2$ text content value;
- $Title$ is the web page title data;
- URL is the Uniform Resource Locator (URL) of a web page that should be parsed.

The Mathematical Model of the EA Web Mining Results

Here:

- V is the set of vertices that represent EA model elements;
- $E \subset V \times V$ is the set of edges that represent relationships between EA model elements;
- C is the set of ArchiMate element types;
- R is the set of ArchiMate relationship types;
- $vt: V \to C$ is the mapping between ArchiMate element types and graph vertices;
- $et: E \to R$ is the mapping between ArchiMate relationship types and graph edges.

$$AM = Build(URL, Title, Desc, BA) =$$

$$= \Big\langle V = \{URL, Desc, Title, "Web\ hosting"\} \cup Text_2 \cup \bigcup_{text_2 \in Text_2} BA(text_2),$$

$$E = \bigcup_{text_2 \in Text_2} \{text_2, Title\} \cup \bigcup_{text_2 \in Text_2} \bigcup_{attr("href") \in BA(text_2)} \{attr("href"), text_2\} \cup$$

$$\cup \bigcup_{text_2 \in Text_2} \bigcup_{attr("href") \in BA(text_2)} \{Title, attr("href")\} \cup \{\{URL, Title\}\} \cup$$

$$\cup \{\{"Web\ hosting", Title\}\},$$

$$C = \{BusinessService, BusinessProcess, ApplicationService, ApplicationComponent, TechnologyService, TechnologyNode\},$$

$$R = \{Realization, Serving, Composition\},$$

$$vt = \{(URL, TechnologyService), (Desc, BusinessService), (Title, ApplicationComponent), ("Web\ hosting", TechnologyNode)\} \cup$$

$$\cup \bigcup_{text_2 \in Text_2} \{(text_2, BusinessProcess)\} \cup$$

$$\cup \bigcup_{text_2 \in Text_2} \bigcup_{attr("href") \in BA(text_2)} \{(attr("href"), ApplicationService)\},$$

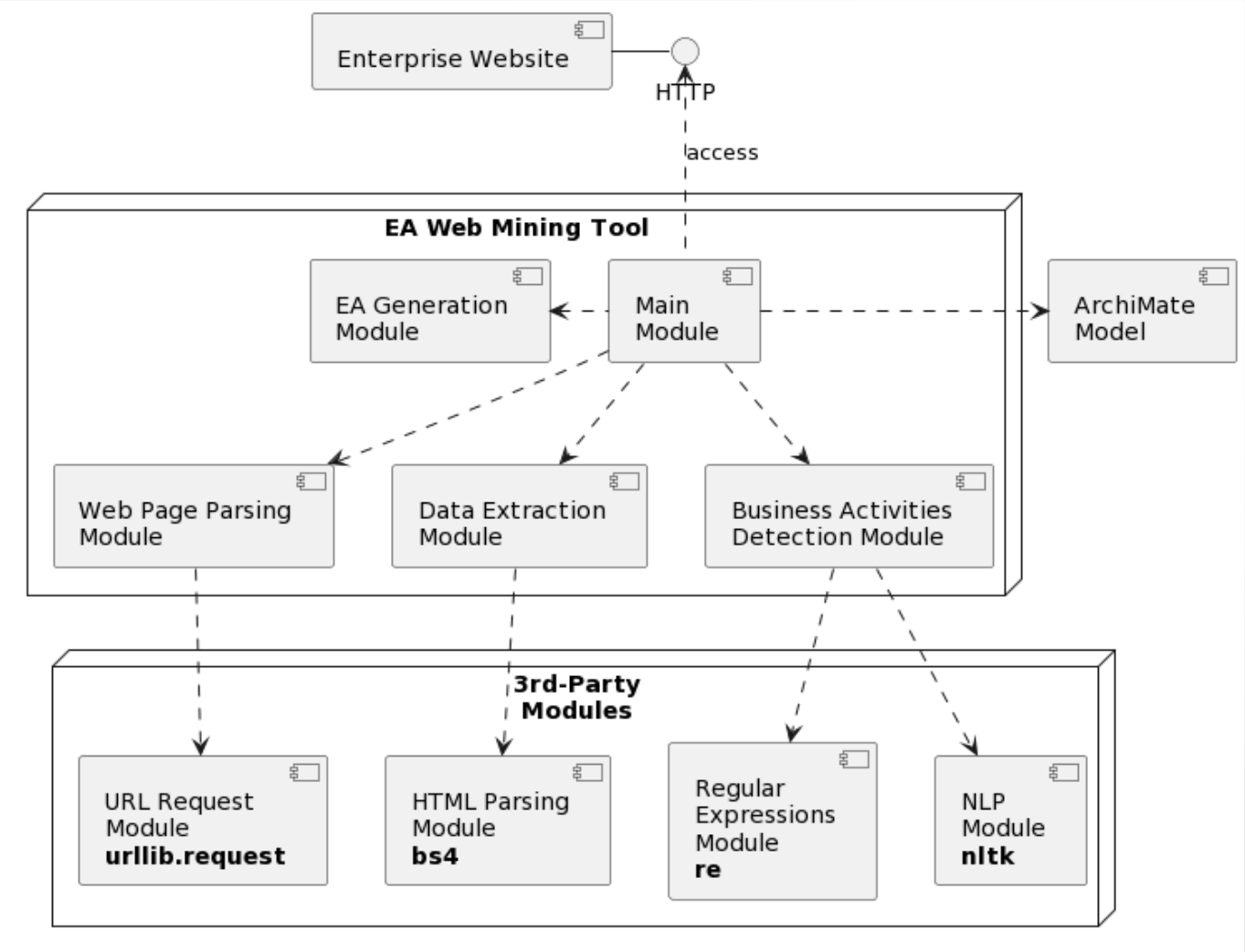
$$et = \{(\{URL, Title\}, Serving)\} \cup \{(\{"Web\ hosting", Title\}, Composition)\} \cup$$

$$\cup \bigcup_{text_2 \in Text_2} \{(\{text_2, Desc\}, Realization)\} \cup$$

$$\cup \bigcup_{text_2 \in Text_2} \bigcup_{attr("href") \in BA(text_2)} \{(\{attr("href"), text_2\}, Serving)\} \cup$$

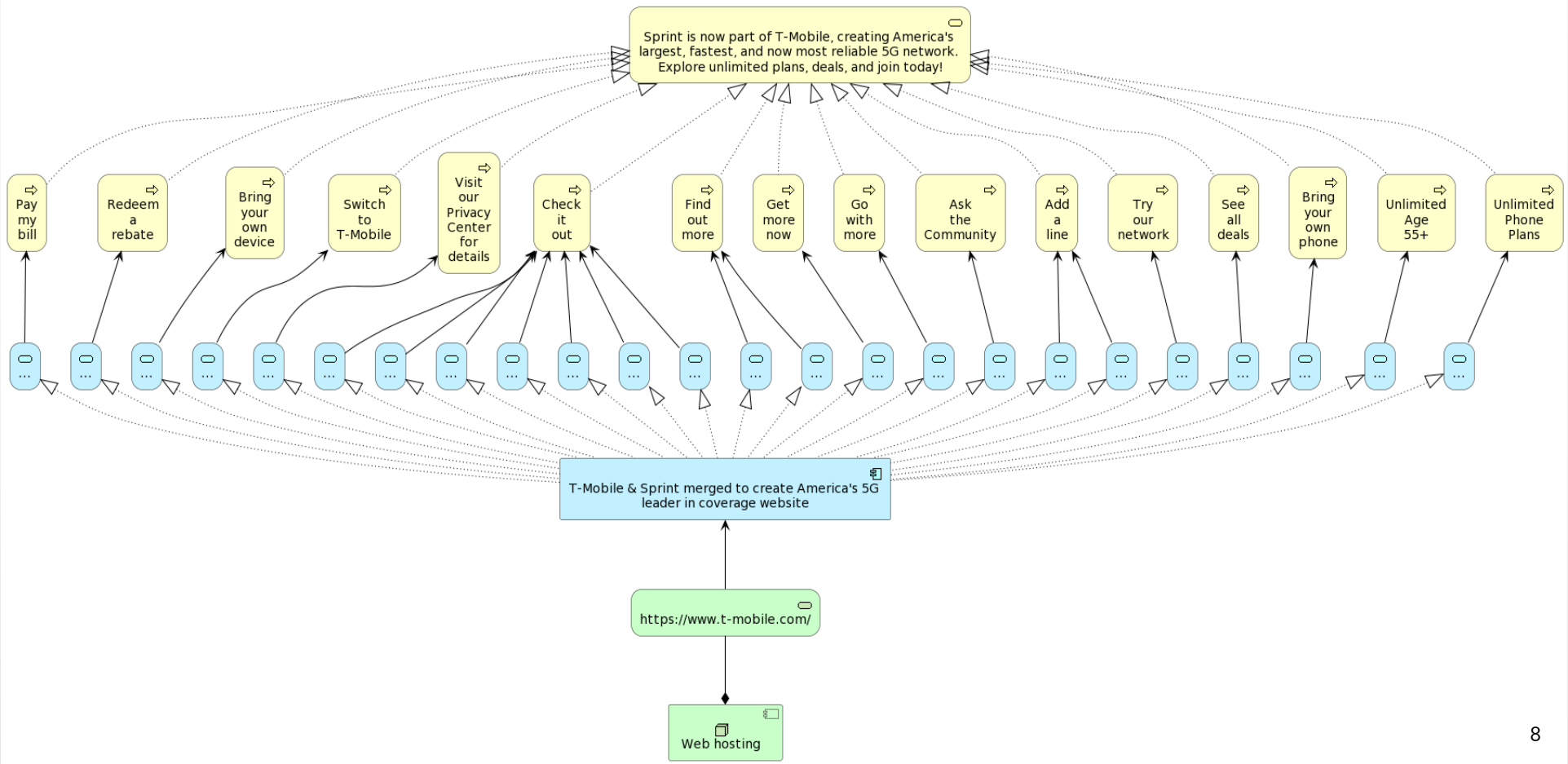
$$\cup \bigcup_{text_2 \in Text_2} \bigcup_{attr("href") \in BA(text_2)} \{(\{Title, attr("href")\}, Realization)\} \Big\rangle$$

Software Implementation of the EA Web Mining Approach

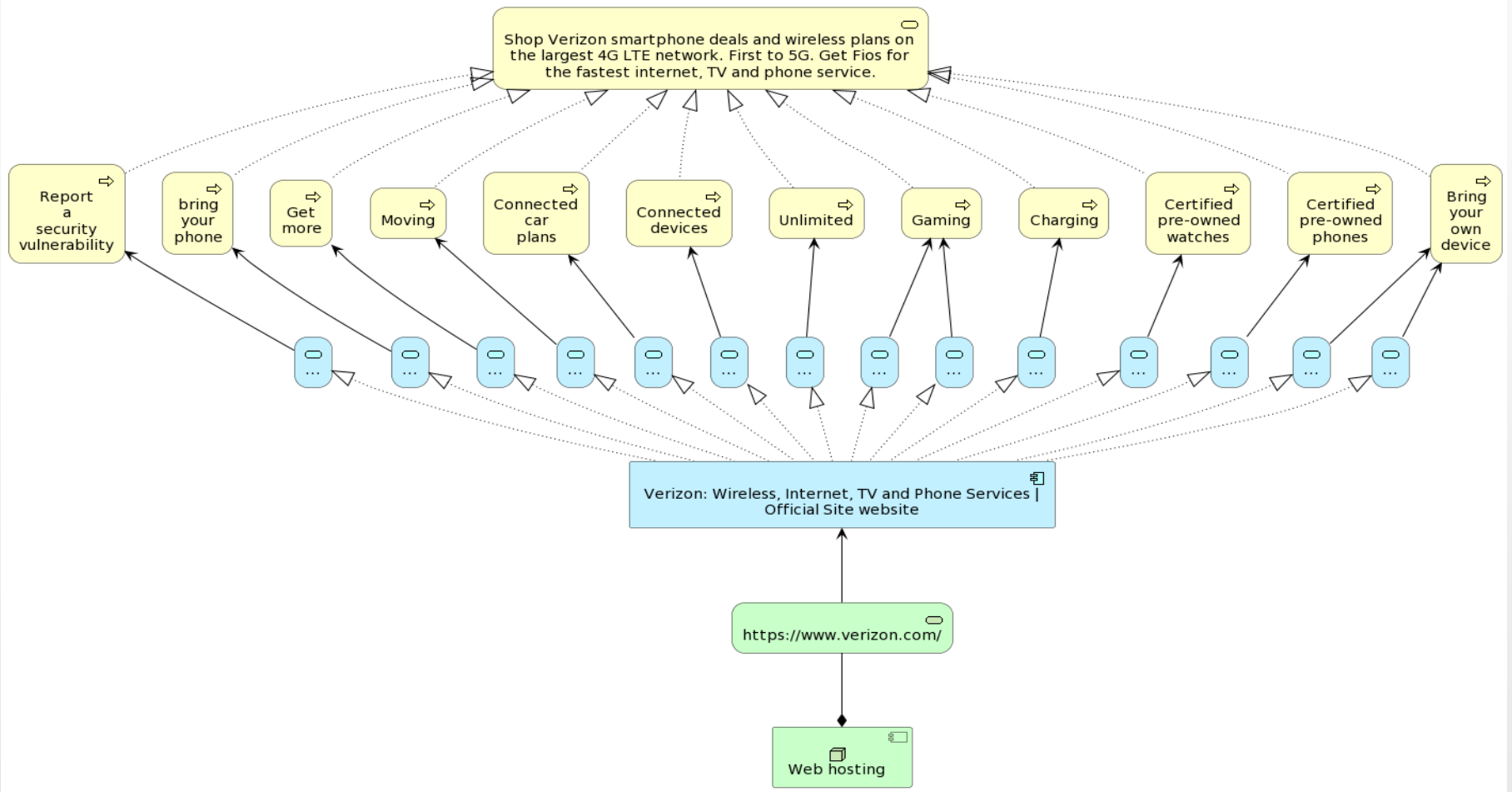


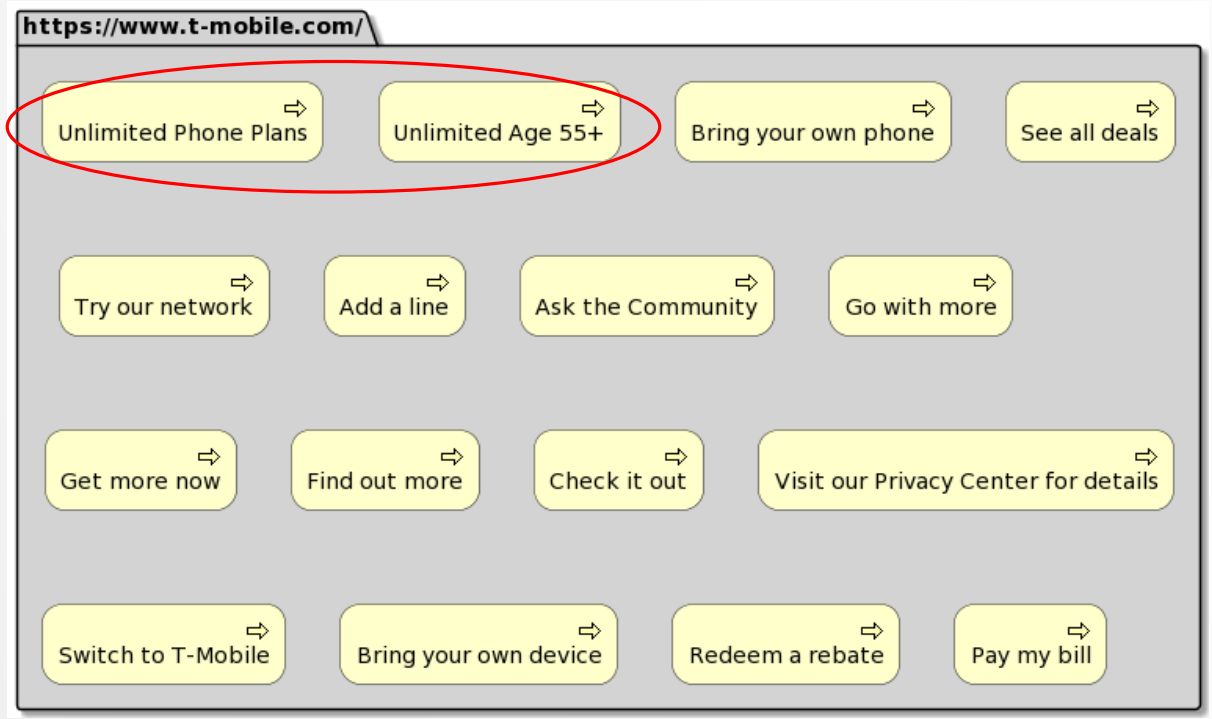
- "**Web Page Parsing**" – HTML page parsing to work with tags, attributes, and text contents;
- "**Data Extraction**" – the title and description tags processing, as well as URL address and text content data extraction from web page hyperlinks;
- "**Business Activities Detection**" – hyperlinks processing to detect the ones that mean certain business activities that trigger business processes supported by the web application services;
- "**EA Generation**" – ArchiMate model generation using EA elements and relationships formulated on the previous steps.

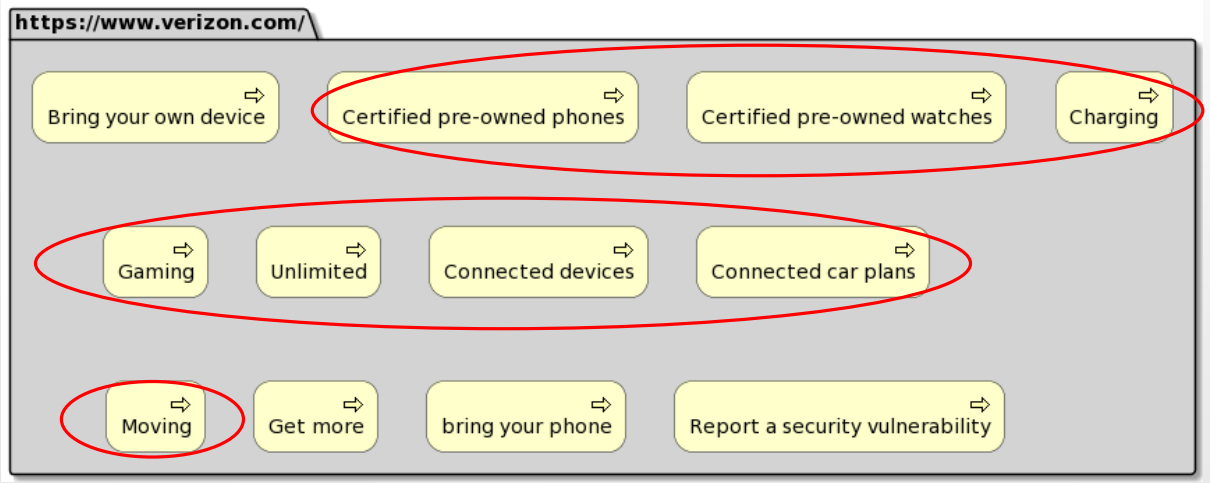
Obtained Results: The EA model built as the result of T-Mobile homepage processing



Obtained Results: The EA model built as the result of Verizon homepage processing







Results Analysis and Discussion

Created **Business Architecture** models demonstrate only the business process architecture, while other EA elements and relationships are avoided.

There are "false positive" business processes that do not correspond to the verb-object activity labeling style:

$$Precision = \frac{TP}{TP + FP} = \frac{14 + 4}{(14 + 4) + (2 + 8)} = \frac{18}{28} = 0.64.$$

The 64% of detected business process elements are representing business activities offered by the considered websites.

Conclusion and Future Work

- The proposed technique is named "**enterprise architecture web mining**" and aims to **simplification of the process of enterprise architecture blueprinting** in the early stages of EA development.
- It is expected that **the proposed approach can reduce the time and cost consumption** of EA modeling by making it possible to construct business process-centric EA landscapes directly from company homepages.
- The proposed approach uses **HTML parsing techniques to extract data** from enterprise web pages.
- Then **detected business activities are represented as ArchiMate business processes** together with remaining EA elements, such as the **business service** (based on the web page description), **application services** (based on the hyperlink URL values), the **application component** (based on the web page title), the **technology service** (based on the web page URL), and the **technology node** (it represents a web hosting).
- Obtained EA models and their analysis results demonstrate **the 64% precision of the suggested "EA mining" technique**.
- Future work in this field should include **the elaboration of business activity detection** in enterprise web pages.

Thank You for Attention!
Questions?